Research Statement

Trevor Campbell

The big data phenomenon continues unabated, with datasets continually growing in size, complexity, and diversity. Alongside this explosion in data has come the redefinition of the "statistics practitioner"; no longer just those with in-depth training in statistics, practitioners arise more and more from other technical disciplines as their main challenges shift from generating data to analyzing data (Nagy, 2016). Scientists, engineers, and analysts are now faced with immense datasets generated by particle accelerators, social networks, cell transcriptomes, financial markets, and more. *Bayesian methods* have a compelling set of advantages in these modern data analysis settings: they provide rich hierarchies that encode complex latent relationships and share information intelligently across subpopulations, posterior distributions with a wealth of options for principled parameter estimation and uncertainty quantification, and priors that enable the incorporation of expert knowledge. But to keep pace with the growth in both datasets and diversity of statistics practitioners, they must also be scalable, easy to implement and tune, and provide reliably good results. Moreover, they must be well-suited to the streaming and distributed data modalities becoming increasingly prevalent.

Researchers have made significant strides in these areas already, leading to many successful uses of Bayesian methods in modern applications. However, there is still significant room for improvement. From an inferential standpoint, Markov chain Monte Carlo methods have rigorous theoretical guarantees, but can be too slow to be practical on large datasets. Variational Bayes and related approaches are often scalable, but lack posterior approximation guarantees and can destroy important posterior structure. Many of these methods are also difficult to implement and require significant tuning effort, making them impractical for nonexpert use. From a modeling standpoint, Bayesian nonparametrics provides flexible models that can grow over time and are naturally suited to streaming data; but scalable inference methods assume the data exists in a single large batch, undermining the primary benefit of nonparametrics.

The fundamental goal of my research is to develop effective, practical, and easy-to-use Bayesian methods for large-scale and streaming data. In particular, my research has two high-level threads: I develop automated, scalable Bayesian posterior approximations with theoretical guarantees, making Bayes accessible to the nonexpert; and I develop Bayesian models and posterior inference algorithms for streaming data, taking advantage of the flexibility of Bayesian nonparametrics.

Automated Bayesian approximation with guarantees

One of the most important recent developments in the Bayesian paradigm has been the shift towards *automation*: rather than having to develop, code, and tune model-specific algorithms, practitioners now have "black-box" implementations that require only a basic specification of the model as inputs. This level of automation enables experts and nonexperts alike to use more sophisticated models, facilitates faster exploratory modeling and analysis, and helps ensure experimental reproducibility.

Bayesian coresets Past work on scaling Bayesian inference has largely focused on making model-specific modifications to standard algorithms, which require expert tuning, can break theoretical quality guarantees, and are not easily automated. My research (Huggins et al., 2016; Campbell and Broderick, 2017) has instead leveraged the redundancy of data in large datasets to obtain a small, weighted subset of the data (called a *Bayesian coreset*) that can be used in place of the full dataset in a standard posterior inference algorithm. Coresets provide scalable inference that is naturally suited to streaming and parallel data modalities. In the initial study (Huggins et al., 2016), collaborators and I provided a coreset construction algorithm based on nonuniform random sampling, along with a guarantee on the worst-case deviation of the coreset log-likelihood from that of the full dataset. While simple and computationally inexpensive, this algorithm is difficult to automate: its implementation requires computing the *sensitivity* of each data point, a model-specific task that involves significant technical expertise. Its performance is also limited by the fact that it cannot use the current coreset to influence the selection of future points.

More recently, I developed a novel perspective of coreset construction as sparse vector sum approximation by reformulating the log-likelihood functions as vectors in a weighted-supremum-norm vector space. Noting that the weighted supremum norm was the bottleneck for both performance and automation, I replaced

it with a norm corresponding to an inner product, resulting in *Hilbert coresets* (Campbell and Broderick, 2017). Hilbert coresets have simple, efficient, and theoretically sound constructions based on the Frank-Wolfe algorithm (Frank and Wolfe, 1956), with geometric convergence guarantees that significantly improve upon the inverse square root guarantees of previous coreset constructions. In addition, they are easily automated due the use of model-agnostic random finite projections (Rahimi and Recht, 2007) for norm approximation. Experiments on a wide variety of models (logistic regression, Poisson regression, directional clustering) and datasets (chemical reactivities, phishing attempts, airport delays, bikeshare rides, 3D indoor environment surface normals, protein backbone configurations) show that Hilbert coresets provide high quality posterior approximations and reduce the computational cost of inference by orders of magnitude.

Truncated random measures Scalable inference algorithms for Bayesian nonparametric models typically require a finite approximation of the infinite-dimensional prior. However, developing such a finite approximation is a technically involved process, requiring detailed knowledge of probability theory. To address this challenge, my research has provided a standardized, black-box methodology for the development of finite approximations of *(normalized) completely random measures ((N)CRMs)*—a rich source of Bayesian nonparametric priors such as the beta, gamma, and Dirichlet processes—through *truncation* of *sequential representations* (Campbell et al., 2016b). In particular, my work detailed two major classes of sequential (N)CRM representations that can be used for simulation and inference—*series representations* and *superposition representations*—along with generalized truncation error and computational complexity analyses for each class. Within the two classes, my work provided numerous different general representations and guidance on which is most appropriate in different scenarios. These provide practitioners with the means to easily generate finite (N)CRM approximations and to set truncation levels based on our error bounds. I have applied the truncations from this work to develop flexible uncertainty sets in robust optimization (Campbell and How, 2015) and to provide a statistically motivated solution to the global point cloud alignment problem in computer vision (Straub et al., 2017).

Bayesian models and inference for streaming data

Streaming data presents even greater challenges for modeling and inference: models must correctly adapt their complexity as more data are observed, and inference algorithms can only examine data at most a small number of times before it must be discarded. My research addresses both of these challenges.

Exchangeable trait allocations Bayesian nonparametrics provides a wealth of models for streaming combinatorial data. Two key questions when using such models are whether their complexity (number of clusters, features, topics, etc.) grows at a rate suitable for the data at hand, and whether efficient inference is possible. Remarkably, in the case of *exchangeable* data—roughly, where the order of observations is irrelevant for inference—these questions have been addressed for *all* clustering and feature allocation models (Kingman, 1978; Broderick et al., 2013b). But clustering and feature allocation do not capture the full spectrum of streaming combinatorial data, such as documents spanning multiple topics in a growing corpus, or edges in a growing social network such as Twitter or Facebook. My work (Campbell et al., 2016a) provides the first characterization of *exchangeable trait allocations*, which allow data to exhibit arbitrary integer membership in multiple *traits*, generalizing a wide range of combinatorial structures. I develop a *paintbox representation* that relates the number of observed traits to the amount of data, and a correspondence between two subclasses of exchangeable trait allocations with the potential for efficient MCMC and variational inference. I also develop the first direct connection between probability functions for clustering, feature allocations, and networks. I have used this general theory to characterize the distribution of all *edge-exchangeable networks* (Cai et al., 2016), a model recently developed by myself and collaborators for realistic sparse graph sequences.

Streaming, distributed variational inference for unsupervised models Bayesian models in which the data are members of unobserved, unordered traits—like exchangeable trait allocations—exhibit *posterior symmetry*, i.e., posterior probabilities are invariant to reordering the traits. In streaming, distributed settings, this poses a problem: since inference algorithms tend to destroy the symmetry structure in the posterior for each data minibatch, merging minibatch posteriors directly (Broderick et al., 2013a) risks incorrectly merging information about different traits. My research has provided a solution for both parametric (Campbell and

How, 2014) and nonparametric (Campbell et al., 2015) unsupervised models, by formulating a combinatorial matching problem between components and providing an efficient optimization algorithm. In the nonparametric case, the prior regularizes the component matching. In contrast to previous algorithms, the proposed framework is streaming, distributed, asynchronous, learning-rate-free, and truncation-free. Experiments on both mixture and topic models on 20-newsgroups, MNIST, SUN images, and an aircraft trajectories dataset show that the proposed algorithms provide orders of magnitude reductions in computation time versus state-of-the-art algorithms, with comparable posterior approximation quality.

Small-variance dynamic clustering In practice, many streams are not exchangeable; for example, exchangeability does not hold if the set of latent traits evolves over time. My work has also provided tractable streaming inference in this setting for nonparametric clustering models with a Markov chain dependence structure (Campbell et al., 2013). Noting that inference on the full Bayesian model is too computationally costly, I used *small-variance analysis*—in which the posterior is connected to a hard clustering problem by considering its limit as component likelihood variances tend to 0—to develop a k-means-like cost function and corresponding alternating minimization algorithm. The clustering algorithm exhibits the scalability typical of k-means-like algorithms, handles cluster birth, death, and motion, and naturally maintains correct cluster correspondences over time. I have applied this technique to cluster real-time streams of computer vision data (Straub et al., 2015) on a laptop with data rates of over 9,000,000 data points per second, and extended the method to kernelized clustering based on arbitrary similarity functions (Campbell et al., 2017).

Future work

Bayesian methods have the potential to tackle some of the most challenging modern data analysis problems. My work on Bayesian methods will continue to push the boundaries of scalability, automation, and theory; and it will be guided by the overarching goal of making Bayes both accessible to the growing diversity of statistics practitioners and applicable to the growing diversity of large-scale, streaming data analysis problems. In the short term, there are numerous specific opportunities to build upon my past work. Hilbert coresets could be advanced by making them applicable to high-dimensional models, exploring the connection of proposed norms to well-known measures of posterior discrepancy, investigating novel automated inner product approximations, obtaining tighter bounds on the quality of random projections, and evaluating more efficient Frank-Wolfe variants (Lacoste-Julien and Jaggi, 2015). Further, my work on general truncated sequential representations leaves open the question of automated inference schemes, which are critical for making these approximate models broadly accessible. Finally, insights from my work on exchangeable trait allocations could be used to develop representations of more general nonexchangeable data streams.

References

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. (2013a). "Streaming variational Bayes." In *Advances in Neural Information Processing Systems*.

Broderick, T., Pitman, J., and Jordan, M. (2013b). "Feature allocations, probability functions, and paintboxes." *Bayesian Analysis*, 8(4): 801–836.

Cai, D., Campbell, T., and Broderick, T. (2016). "Edge-exchangeable graphs and sparsity." In *Advances in Neural Information Processing Systems*.

Campbell, T. and Broderick, T. (2017). "Automated scalable Bayesian inference via Hilbert coresets." *arXiv:1710.05053*.

Campbell, T., Cai, D., and Broderick, T. (2016a). "Exchangeable trait allocations." *arXiv:1609.09147*.

Campbell, T. and How, J. P. (2014). "Approximate decentralized Bayesian inference." In *Uncertainty in Artificial Intelligence*.

— (2015). "Bayesian nonparametric set construction for robust optimization." In *American Control Conference*.

Campbell, T., Huggins, J., How, J., and Broderick, T. (2016b). "Truncated random measures." *arXiv:1603.00861*.

Campbell, T., Kulis, B., and How, J. (2017). "Dynamic clustering algorithms via small-variance analysis of Markov chain mixture models." *arXiv:1707.08493*.

Campbell, T., Liu, M., Kulis, B., How, J. P., and Carin, L. (2013). "Dynamic clustering via asymptotics of the dependent Dirichlet process mixture." In *Advances in Neural Information Processing Systems*.

Campbell, T., Straub, J., Fisher III, J. W., and How, J. P. (2015). "Streaming, distributed variational inference for Bayesian nonparametrics." In *Advances in Neural Information Processing Systems*.

Frank, M. and Wolfe, P. (1956). "An algorithm for quadratic programming." *Naval Research Logistics Quarterly*, 3: 95–110.

Huggins, J., Campbell, T., and Broderick, T. (2016). "Coresets for scalable Bayesian logistic regression." In *Advances in Neural Information Processing Systems*.

Kingman, J. F. C. (1978). "The representation of partition structures." *Journal of the London Mathematical Society*, 2(2): 374–380.

Lacoste-Julien, S. and Jaggi, M. (2015). "On the global linear convergence of Frank-Wolfe optimization variants." In *Advances in Neural Information Processing Systems*.

Nagy, R. (2016). "The era of big data is coming: scientists need to step out of their comfort zone." Posted by Jack Leeming: http://blogs.nature.com/naturejobs/2016/09/26/the-era-of-big-data-is-coming-scientists-need-to-step-out-of-their-comfort-zone/.

Rahimi, A. and Recht, B. (2007). "Random features for large-scale kernel machines." In *Advances in Neural Information Processing Systems*.

Straub, J., Campbell, T., How, J. P., and Fisher III, J. W. (2015). "Small-variance nonparametric clustering on the hypersphere." In *IEEE Conference on Computer Vision and Pattern Recognition*.

— (2017). "Efficient global point cloud alignment using Bayesian nonparametric mixtures." In *IEEE Conference on Computer Vision and Pattern Recognition*.